Judging spatial relevance and document location for Geographic Information Retrieval

Purves, Ross S ; Clough, Paul D

Judging spatial relevance and document location for Geographic Information Retrieval

Ross S. Purves[1] and Paul D. Clough[2]

[1]Department of Geography, University of Zurich
rsp@geo.unizh.ch

[2]Department of Information Studies, University of Sheffield
p.d.clough@sheffield.ac.uk

INTRODUCTION

Finding documents relevant to queries which include some geographical context is key to successful Geographic Information Retrieval (Larson, 1996). Such queries typically consist of a triplet containing a theme, spatial relationship and location, for example <castles><near><Edinburgh>. SPIRIT (Jones et al., 2004) is a prototype spatially-aware search engine designed to deal with such queries by considering not only the textual content of documents, but also assigning multiple geographic footprints to all toponyms identified within document text (Clough, 2005).

Document retrieval is based on a query footprint calculated by a geographic ontology (Fu et al., 2005) from a footprint based on location (e.g. Edinburgh) and a spatial relationship (e.g. near, inside, outside, north of, west of, etc…). Documents are retrieved and ranked using a combination of traditional text ranking methods and geometric calculations based on footprints. We have previously argued that evaluation of GIR must consider both thematic and spatial relevance but found that evaluators experienced difficulties in evaluating spatial relevance, and we suggested that this was in part due to a need for local knowledge of a region to evaluate a document's spatial relevance (Clough et al., 2006).

In this paper we focus on two components of spatial relevance. Firstly, we assess for the top 10 documents retrieved for 20 queries the spatial relevance of *content locations* with respect to the spatial component of the query (i.e. "beaches in Cornwall"). Wang et al. (2005) defined content location as being the geographic area a web page refers to. Secondly, we judge whether the footprints retrieved are relevant to the content location of the document with which they are associated.

METHODOLOGY

We performed 20 queries using SPIRIT which illustrate a range of spatial relationships and granularities with respect to query footprint (Tab. 1).

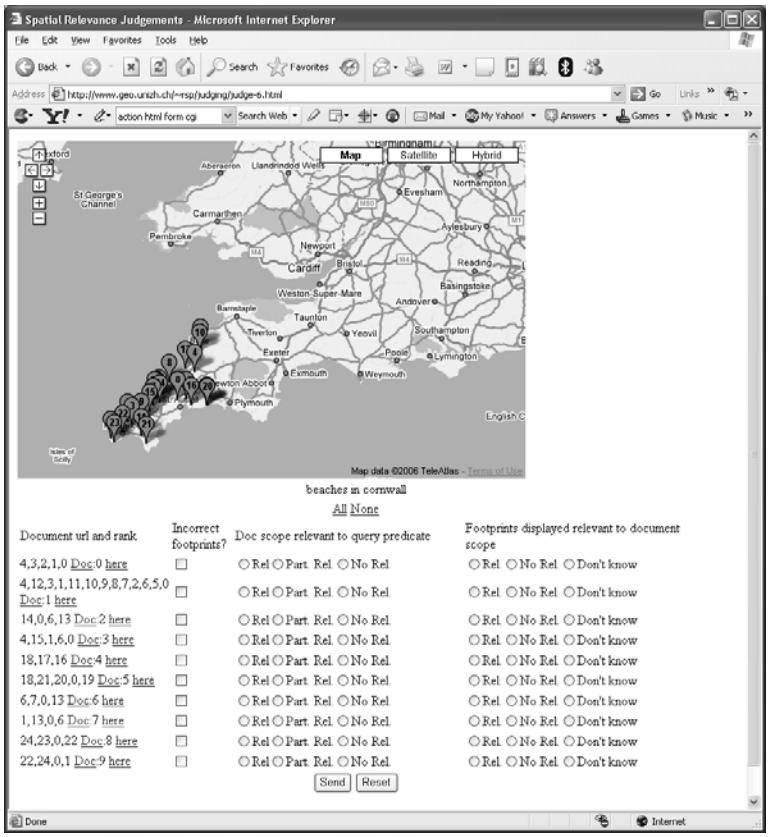

Fig. 1: Relevance judging application for query <beaches><inside><cornwall>

Query results, consisting of document URLs and 1…n footprints associated with the document and relevant to the query, were displayed in a simple application utilising the Google Maps API (Fig. 1) which allowed viewing of footprint centroids and content associated with individual documents. For each document we judged the relevance of a document's content location with respect to a query (e.g. a document describing multiple museums in Wales would not be relevant to a query for <museums><inside><Cardiff>, since Wales in not inside Cardiff) and the relevance of the footprints displayed with respect to content location (e.g. for the same document, a footprint in Cardiff is relevant to the document's location of Wales since Cardiff is inside Wales).

RESULTS AND DISCUSSION

The results illustrated that, in general, spatial search performed well with mean precision for the 20 queries of 0.51 (Tab. 1). The queries with the worst performance tended to be those where the granularity of the query was consumerate with the finest level of granularity in the ontology and thus few documents whose content location exactly matched the query location were found. There seemed to be no strong relationship between precision and the nature of the spatial relationship.

However, from this preliminary study we believe the most interesting result relates to the inter-annotator agreement. The interannotator agreement for spatial relevance is strongly related to the nature of the query type, with much higher agreement values in relevance judgements for queries of type inside – for a 1-tailed Mann-Whitney U test we found a significant relationship ($U=13.5$; $p_0 < 0.01$). Since footprint judgement with respect to document location is independent of query this ordering of agreement should not persist – and indeed we find no significant correlation in ordering ($U=37.5$; $p_0 > 0.1$).

These preliminary results show, firstly, that spatial search seems to perform well for a range of query types and locations. Secondly, and we believe importantly, they show that judging the relevance of non-containment type operators appears to be more variable and thus an important element not only in evaluating GIR but also in consideration of the development of algorithms aiming to represent subjective qualities such as *nearness* in GIR systems (Worboys, 2001). Further work will address a similar range of queries of differing granularities with more annotators.

Tab. 1: Summary of experimental results showing P10 for each query and mean number of footprints relevant to document scope for displayed footprints. Non-containment queries are shaded in Table and entries are ordered by agreement in spatial relevance judgements

Query	Mean P10	Agreement in spatial relevance judgements	Mean number of footprints falling within document location	Agreement in document location judgements
beaches in cornwall	1.00	0.90	0.90	0.80
camping in highland	0.55	0.89	0.55	1.00
museums in cardiff	0.35	0.80	0.55	1.00
red kites near cromarty	0.10	0.80	0.05	0.88
pubs in edinburgh	0.55	0.80	0.55	0.80
shipping in liverpool	0.40	0.80	0.45	0.56
hotels in cardiff	0.25	0.70	0.75	0.50
schools in norwich	0.50	0.67	0.50	0.67
mountaineering in scotland	1.00	0.60	0.80	1.00
castles east of edinburgh	0.60	0.60	0.60	0.60
hotels near edale	0.30	0.60	0.60	0.56
walking in fife	0.44	0.56	0.40	0.56
camping near lancaster	0.40	0.56	0.60	0.44
oil industry in aberdeen	0.50	0.50	0.40	0.57
hotels west of fort william	0.42	0.50	0.75	0.50
canals near stroud	0.40	0.40	0.65	0.63
cycling south of london	0.40	0.40	0.30	0.43
walking outside of edinburgh	0.59	0.38	0.35	0.83
climbing near aviemore	0.85	0.33	0.65	0.80
walking near beauly	0.55	0.10	0.80	0.86
Mean values	0.51	0.59	0.56	0.70

ACKNOWLEDGEMENTS

This research was supported by the EU-IST Project No. IST-2001-35047 (SPIRIT) and the Swiss BBW (01.0501).

REFERENCES

Clough, P. (2005) Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. Proceedings of the 2005 Workshop on Geographic Information Retrieval. Bremen, Germany.

Clough, P. Purves, R.S. & H. Joho, H. (2006) Judging the spatial relevance of documents for GIR. Lecture Notes in Computer Science, 3936.

Larson, R., 1996, "Geographic Information Retrieval and Spatial Browsing." In Smith, L. & Gluck, M. (eds.) *GIS and Libraries: Patrons, Maps and Spatial Information.* University of Illinois.

Jones, C. B., A.I. Abdelmoty, et al. (2004) The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Lecture Notes in Computer Science, 3234.

Fu, G., C.B. Jones & A.I. Abdelmoty (2005) Building a Geographical Ontology for Intelligent Spatial Retrieval on the Web. Proceedings of IASTED International Conference on Databases and Applications (DBA2005). Innsbruck, Austria.

Wang, C., Xie, X. et al. (2005) Detecting geographic locations from web resources**.** Proceedings of the 2005 Workshop on Geographic Information Retrieval. Bremen, Germany.

Worboys, M. F. (2001) Nearness relations in environmental space. International Journal of Geographical Information Science, 15, 7, 633 – 651.